

Estymacja w modelu logistycznym z częściową obserwowalnością typu PU

Jan Mielniczuk

Instytut Podstaw Informatyki PAN i Wydział
Matematyki i Nauk Informacyjnych PW

W referacie rozważymy problem estymacji w modelu logistycznym dla pary zmiennych $(X, Y) \in R^p \times \{0, 1\}$, gdzie zmienna binarna Y jest nieobserwowalna i jest zastąpiona przez obserwowalną zmienną binarną S , przy czym $P(S = 1|Y = 1, X) = c < 1$ oraz $P(S = 0|Y = 0, X) = 1$ (model PU; *Positive Unlabelled*). Przedyskutowane zostanie metoda naiwna, polegająca na dopasowaniu modelu logistycznego do obserwacji z rozkładu $P_{X,S}$ oraz metoda największej wiarygodności dla jednoczesnej estymacji parametrów β modelu logistycznego i częstości etykietowania c . W szczególności pokazane zostaną związki metody naiwnej ze złą specyfikacją i własność ściągania parametru β w kierunku 0. Przedstawione zostaną również wyniki badań symulacyjnych dla estymacji prawdopodobieństwa a posteriori i c . Wyniki uzyskano wspólnie z M. Łazęcką i P. Teisseyrem.

Literatura

- [1] J. Baker, J. Davis (2020), *Learning from positive and unlabelled data: a survey*, Machine Learning, 109, 719-760
- [2] P. Teisseyre, J. Mielniczuk, M.Łazęcka (2020), *Different strategies for fitting logistic regression for positive and unlabelled data*, Proceedings of the International Conference on Computational Science ICCS 2020, LNCS, 3-17